STAC67: Regression Analysis

Assignment 2
(Total: 100 points)

Please submit R Markdown file for Q. 1- Q. 2 along with your submission of the assignment.

Q. 1 (20 pts) This is to practice R for simulations.

When you generate a random number, use R code, **set.seed(your student number)** before the R codes of generating a random number, so that we can replicate the result.

We start by assuming true regression parameters in the model. Thus, we assume that $Y_i = 46.3 + 4X_i + \epsilon_i$, with $\epsilon_i \sim N(0, 3.9^2)$. We use the predictors X (growth) that we already have from "vote.txt".

- Step 1: Simulation of the fake data
  Simulate a vector Y of fake data and put this in a data frame with the same X (growth).
- Step 2: Fitting the model and keeping the estimated regression coefficients.
- Step 3: Repeating Step 1 and Step 2, 10,000 times.

(a) (5 pts) Do Step 1 and Step 2. Obtain the least square estimates of β_0 and β_1 with the fake data.
Also, compute estimated $E(Y|X_0 = 0.1)$ and obtain 95% confidence interval for $E(Y|X_0 = 0.1)$ by hands and compare it by R built-in function.

(b) (10 pts) Do Step 3. Make a histogram of 10,000 $\hat{\beta}_0$ and 10,000 $\hat{\beta}_1$. Superimpose (overlay) its theoretical distribution on each histogram. Calculate the mean and standard deviation of 10,000 estimates each. Are the results consistent with theoretical values?

(c) (5 pts) Do Step 3. Generate 10,000, 95% confidence interval for $E(Y|X_0 = 0.1)$. What proportion of the 10,000 confidence intervals for $E(Y|X = 0.1)$ includes $E(Y|X = 0.1)$? Is this result consistent with theoretical expressions?

Q. 2 (40 points) The dataset "NBAhtwt.csv" is posted at Quercus. It contains weights (Y, in pounds) and heights (X, in inches) for 505 National Basketball players for the 2013/2014 season. Complete the following parts (treating this as a sample from a conceptual population of potential athletes).

(a) (5 pts) Fit a Simple Linear Regression relating Weight (Y) to height (X) using R. Construct 95 % confidence interval for the mean weight of all players with $X_0 = 74$. Compute it by hands (use R) and compare the result with the built-in R function.

(b) (5 pts) Construct a 95% prediction interval for a new player with $X_0 = 74$. Compute it by hands (use R) and compare the result with the built-in R function.

(c) (5 pts) Construct the Analysis of Variance table, and interpret the R^2.

(d) (5 pts) Plot the residuals versus fitted values. Comment on residual plot.

(e) (5 pts) Obtain a normal probability plot of residuals and test the hypothesis that the errors are normally distributed with the Shapiro-Wilk test.

(f) (10 pts) We would like to conduct the **Brown-Forsythe test** to determine whether or not the error variance varies with the level of X. Divide the data into the two groups based on the median of X. Use $\alpha = 0.05$. Do not use the built-in R function. Write your own R function to implement this test. What is your test result?

(g) (5 pts) If there is evidence of non-normality or non-constant variance of errors, obtain a Box-Cox transformation, and repeat the previous parts (e) and (f).

Q. 3 (20 pts) (5 pts each) An experiment is conducted, relating weekly sales for a food delivery (Y) service to the amount of advertising (X) during the week. The results for a sample of n = 6 weeks are given below.

X_i:	2	2	4	4	6	6
Y_i:	20	30	40	50	70	60

Use the simple linear regression in matrix form.

(a) Obtain the design matrix $\mathbf{X}$ and $\underset{\sim}{Y}$.

(b) Obtain the vector of estimated regression coefficients, $\underset{\sim}{\hat{\beta}}$, and the vector of fitted value, $\underset{\sim}{\hat{Y}}$, and the residual vector, $\underset{\sim}{e}$.

(c) Compute the estimated variance-covariance matrix of $\underset{\sim}{\hat{\beta}}$, $\widehat{Var}(\underset{\sim}{\hat{\beta}})$.

(d) Find the hat matrix $\mathbf{H}$. What does $\sum_{i=1}^{n} h_{ii}$ equal? Here, h_{ij} is the element in $\mathbf{H}$ in the ith row and jth column.

(e) Find the estimated variance-covariance matrix of the residual vector, $\widehat{Var}(\underset{\sim}{e})$.

Q. 4 (20 pts) (5 pts each) The total salaries (X, in millions of pounds) and the number of points earned (Y) for the n = 20 English Premier League teams in 1995/6 are used to fit a simple linear regression model. For this problem, we will treat this as a sample from a population of all possible league teams. Here are the matrices that we obtained from this data.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 20 & 170.9 \\ 170.9 & 1745.15 \end{bmatrix} \quad \mathbf{X}'\underset{\sim}{Y} = \begin{bmatrix} 1042 \\ 9870.3 \end{bmatrix}$$

(a) Compute $(X'X)^{-1}$ and $\underset{\sim}{\widehat{\beta}}$.

(b) Given that $\bar{Y} = 52.1$ and $\sum_{i=1}^{20}(Y_i - \bar{Y})^2 = 4367.8$. Construct the ANOVA table based on this information.

(c) Provide 95% confidence interval for β_1.

(d) Test $H_0 : \beta_1 = 0$ vs $\beta_1 \neq 0$ with $\alpha = 0.05$.